Model-based redesign of global transcription regulation.

J. Carrera, G. Rodrigo, A. Jaramillo

Nucleic Acids Research Advance Access published February 2, 2009

Nucleic Acids Research, 2009, 1–11
doi:10.1093/nar/gkp022

Model-based redesign of global transcription regulation

Javier Carrera[1,2], Guillermo Rodrigo[1] and Alfonso Jaramillo[3,4,*]

[1]Instituto de Biologia Molecular y Celular de Plantas, CSIC, [2]Instituto de Aplicaciones en Tecnologias de la Informacion y las Comunicaciones Avanzadas (ITACA), Universidad Politecnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain, [3]Laboratoire de Biochimie, Ecole Polytechnique - CNRS, Route de Saclay, 91128 Palaiseau Cedex and [4]Epigenomics Project, Universite d'Evry Val d'Essonne - Genopole - CNRS, 523 Terrasses de l' Agora, 91034 Evry Cedex, France

Received July 12, 2008; Revised January 2, 2009; Accepted January 7, 2009

ABSTRACT

Synthetic biology aims to the design or redesign of biological systems. In particular, one possible goal could be the rewiring of the transcription regulation network by exchanging the endogenous promoters. To achieve this objective, we have adapted current methods to the inference of a model based on ordinary differential equations that is able to predict the network response after a major change in its topology. Our procedure utilizes microarray data for training. We have experimentally validated our inferred global regulatory model in *Escherichia coli* by predicting transcriptomic profiles under new perturbations. We have also tested our methodology *in silico* by providing accurate predictions of the underlying networks from expression data generated with artificial genomes. In addition, we have shown the predictive power of our methodology by obtaining the gene profile in experimental redesigns of the *E. coli* genome, where rewiring the transcriptional network by means of knockouts of master regulators or by upregulating transcription factors controlled by different promoters. Our approach is compatible with most network inference methods, allowing to explore computationally future genome-wide redesign experiments in synthetic biology.

INTRODUCTION

Molecular regulations govern the cell response under environmental (extracellular) or genetic (intracellular) perturbations. The elucidation of these regulations with computational techniques will allow analyzing the cell behavior (1), since modeling in biology has boosted the understanding of the cell mechanisms by means of systemic approaches (2). On the other hand, the design of new transcriptional networks requires a quantitative description of the transcription regulation. Thanks to the new developments in the inference from transcriptomic data, now it is possible to reconstruct the regulatory network with enough accuracy to predict the gene expression profile in presence of heterologous networks. We propose a procedure that, by extending a recent methodology, could be used to redesign transcriptional networks.

The continuous developments on genome sequencing and annotation allow us to design microarrays and to identify the genes and transcription factors (TFs) of an organism. The development of the microarray technology has provided high-throughput genomic measurements, where cells are subjected to several conditions or stresses to measure their gene expression profiles (3). Large-scale cell models, such as metabolic, transcription or protein networks, are distilled from high-throughput genomic data, which poses one of the most challenging problems in biology. The construction of a deterministic model would allow the prediction of the cell response under different stimuli (4).

To redesign the transcriptional regulation network, we need a quantitative model able to predict the gene dynamics. We propose to characterize such model by using microarray data with a known transcriptional network inference method. We first infer the network topology and we later estimate the corresponding kinetic parameters. For the last decade, there has been an enormous effort in the improvement of techniques aimed at the inference of the connectivity of the transcription network. Clustering approaches (5–9) have been used to obtain information of regulatory networks but with low accuracy (10). Information-theoretic inference provides more accurate networks (11–15) even from reduced expression

*To whom correspondence should be addressed. Tel: +33 1 69474444; Fax: +33 1 69474437; Email: alfonso.jaramillo@polytechnique.fr

The authors wish it to be known that, in their opinion, the first two authors have contributed equally to this work.

datasets. A local significance calculation has been very fruitful to capture the network topology (14). On the other hand, Bayesian methods (16–19) give networks with high precision but low proportion of true recovered interactions (they introduce few regulations with high confidence). Moreover, such methods have a higher computational cost. Herein, we propose the construction of predictable genome models in a standard format from a regulatory scaffold captured by using probabilistic methods. Other approaches, instead, optimized directly the corresponding kinetic parameters for a linear regulatory model (20,21). In addition, recent algorithms (22,23) applied sparse logistic regression (24) for gene selection in order to avoid overfitting.

METHODS

We aim to the development of a methodology able to *in silico* evolve a genome for having a predefined transcriptional profile. For this, we require to construct a predictive genome model of transcription, based on ordinary differential equations (ODEs), to account for global redesigns of the cellular regulatory map. Using such models we could study the evolution of gene regulations as a consequence of the environmental stimuli. To construct this we have to use as input microarray data properly normalized (Figure 1). In general, transcription involves protein–DNA interactions, but microarray data gives the genetic expression by quantifying the amount of mRNA. Thus, inferring just from transcriptomic profiles could introduce some inaccuracies due to, for instance, protein–protein interactions of TFs (25,26). Furthermore, some environmental stresses (e.g. heat shock) can alter globally protein expression. However, in this work we neglect these effects for simplicity, assuming that the mRNA amount is proportional to the protein expression and that it is function of the TFs only. In addition, as the precise kinetic model of transcription regulation is not known for any organism, we have generated *in silico* genomes having random regulatory maps with scale-free topology (27). We have applied our methodology against synthetic transcriptomic profiles. We will only assume a previous knowledge of the list of all genes and TFs obtained from genome annotation [e.g. RegulonDB (28) for *Escherichia coli*]. Eventually, we can consider the genomic organization in operons (especially in case of bacteria). Such operons can be known *a priori* or inferred from the same microarray data. Our approach consists of two nested steps. First, we obtain the topology of the network (i.e. which TF regulates which gene or operon) by using an information theory-based approach. We store in a matrix the likelihood of the mutual information (MI) among all the TFs and operons (29–31), computed as the z-scores from the distribution of MI using the transcriptomic expressions for all the perturbing conditions (14). Then, using a suitable threshold, we infer the TFs regulating a given operon. Subsequently, for each operon we perform a multiple linear regression against the corresponding TFs to recover the model kinetic parameters (32). To infer cooperative regulations, we create a set of artificial TFs whose expression profiles are obtained in a combinatorial way as the product of two TF profiles (with the aim of conserving linearity in the formalism). This model is subsequently exported into a SBML file (33), which could be visualized using Cytoscape (34). We have measured the performance of our algorithm by using synthetic transcriptomic data from artificially generated networks.

Mathematical model

We describe the genetic regulations using a linear model for the mRNA dynamics. Here, we use as input data mRNA expression profiles in steady state derived from transcriptional perturbations. As transcriptomic data is normalized and usually represented in logarithmic scale, we have considered $\log_s$(mRNA) as variables (where s can be 2 or 10). Therefore, the mRNA dynamics from gene y_i is given by

$$\frac{\mathrm{d}}{\mathrm{dt}} y_i = a_i + \sum_{j \in TF} b_{ij} y_j + \sum_{j \in TF} \sum_{k \in TF} b_{ijk} y_j y_k - \delta_i y_i, \qquad 1$$

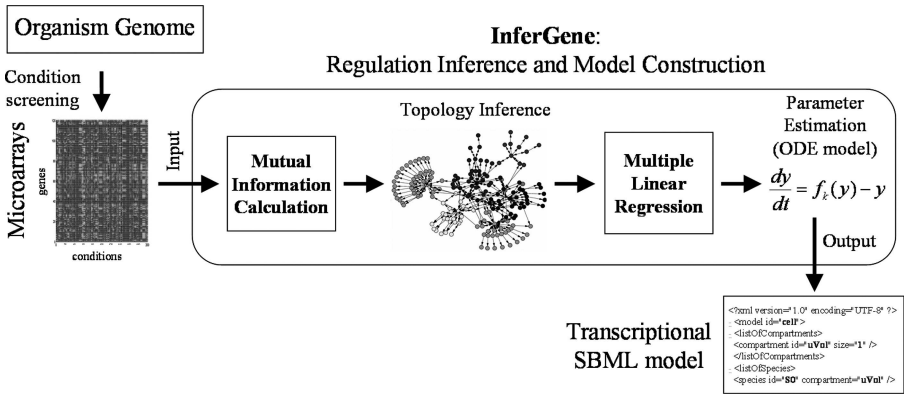

Figure 1. Scheme to infer the regulatory network of an organism. Our inference algorithm uses microarray data and prior knowledge about operons and TFs to predict the full transcriptional regulatory map. It consists of two nested steps: (i) inference of the topology using MI; and (ii) the estimation of the kinetic parameters via multiple linear regressions (Figure S1). We export the constructed model in SBML format (33). We apply our methodology to infer the *E. coli* genome model by using the M3D compendium versus 3 (41) and a list of TFs and operons from RegulonDB (28).

where a_i is the basal synthesis rate, b_{ij} the transcription regulatory coefficient of TF j, b_{ijk} the cooperative transcription regulatory coefficient of TFs j and k acting on the promoter controlling the gene i and δ_i the degradation rate. We set $b_{ij} = 0$ and $b_{ijk} = 0$ when j and j, k are not TFs regulating the gene i. We assume that all the genes of an operon have the same expression value. We also consider that two regulators could act in a cooperative way (i.e. synergistic inductions and cooperative repressions). We do not consider cooperation between more than two TFs.

Here, we use expression values in steady state. Nevertheless, it could be also possible to extend our approach to the use of time series to enrich the experimental input (35). Hence, in the steady state we can write

$$y_i = \alpha_i + \sum_{j \in TF} \beta_{ij} y_j + \sum_{j \in TF} \sum_{k \in TF} \beta_{ijk} y_j y_k, \qquad \mathbf{2}$$

where we have defined $\alpha_i = a_i/\delta_i$, $\beta_{ij} = b_{ij}/\delta_i$ and $\beta_{ijk} = b_{ijk}/\delta_i$. Notice that the resulting parameters are referred to the intensity scale of the microarray technology. We use a time scale such that the mRNA degradation constant is $\delta = 1$. To use a realistic mRNA degradation constant, it would require translating the Affymetrix (36) data to concentration units.

Using network inference to obtain a kinetic model

To obtain a kinetic model suitable for redesign, we take advantage of recent methods aimed to infer the topology of the global regulatory map. In particular, we have chosen one of the best performing methods, the CLR (14), although other methodologies providing a transcriptional map, such as sparse Bayesian methods (19) could also be used. Our approach consists of using multiple regressions to fit the kinetic parameters of a continuous model of the transcription regulation. The approach for large-scale transcription inference is based on measuring the influence between the expression levels of TFs and operons across a large set of conditions. Here, we use MI to estimate the correlation between a TF t and an operon p by using $\mathrm{MI}(y_t, y_p) = H(y_t) + H(y_p) - H(y_t, y_p)$, where H is the entropy of a variable. It is defined as $H(y_i) = -\sum_c p(y_{ic})\log(\mathrm{p}(y_{ic}))$, where y_{ic} is the expression value of gene i in the condition c, and $p(y_{ic})$ the probability to reach that value. The MI is always a positive magnitude. Joint normal distributions are generated with independent variables MI_i and MI_j (values for gene i and TF j, in row i and column j). Thus, the MI matrix is converted into Z matrix where $Z_{ij} = \sqrt{Z_i^2 + Z_j^2}$ and Z_i and Z_j are the z-scores of MI_{ij} from the marginal distributions. According to this matrix, we obtain the genomic interactions.

For completeness, we have developed an algorithm (InferOpe) to infer operons from microarray data. Since two genes from one operon share the same mRNA molecule, we would expect that their transcriptomic profiles would be similar. Our operon prediction is based on the use of co-expression patterns (37), assuming that two genes, i and j, belong to the same operon if they are highly correlated. We evaluate this by using the Pearson correlation coefficient (we assume correlation if $\rho_{ij} > \rho_0 = 0.5$). Moreover, we impose that the angle (θ_{ij}) of such correlation should be around 45° {i.e. $\tan(\theta_{ij}) \simeq 1$], where the relationship with ρ_{ij} is given by $\tan(\theta_{ij}) = \rho_{ij}\frac{\sigma_j}{\sigma_i}$.

For each operon we compute the kinetic parameters for the TFs regulating its promoter. The experimental value of one operon is computed as the average of the expressions of all genes belonging to that operon (i.e. $y_{op} = (1/n)\sum_{g \in op} y_g$, where n is the number of genes of the corresponding operon). To estimate the model parameters α_i, β_{ij} and β_{ijk} we use multiple linear regression (32), which is the result of a minimization problem (least squares) defined by

$$(\widehat{\alpha_i}, \widehat{\beta_{ij}}, \widehat{\beta_{ijk}}) = \arg\min \left\{ (y_i - \alpha_i - \sum_{j \in TF} \beta_{ij} y_j - \sum_{j \in TF} \sum_{k \in TF} \beta_{ijk} y_j y_k)^2 \right\}. \qquad \mathbf{3}$$

We assume that the variability in the experimental conditions and the complexity of the natural regulation is high enough to prevent linear correlations between TFs, which would produce identifiability problems in the regression parameters. Even in such a case, our model is a valid solution although there could be alternative models. We have used the LINPACK libraries (38) to calculate the solution.

Our procedures are implemented in C++, and they run on any UNIX environment. The InferGene software, a tutorial, the corresponding files and some examples are available upon request. The software consists of different functional modules to compute first the network topology and then the corresponding kinetic parameters (see Supplementary Figure S1). Below we present the procedure implemented in InferGene:

(1) Represent the microarray data organized in matrix form, for instance, genes in rows and conditions in columns.
(2) Obtain the list of TFs for the given organism.
(3) Ensure that the microarray matrix contains the expression profiles for all TFs.
(4) Add new rows corresponding to the combinations of two TFs obtained as the product of them (i.e. $y_{TF_i} \cdot y_{TF_j}$ are the new TF profiles).
(5) In case of bacteria, have a file containing the list of operons with the corresponding genes. Otherwise, run InferOpe, our algorithm to infer clustered genes based on co-expression patterns. To maintain the same scheme in all cellular contexts, we can dispose one gene per operon in case of eukaryotes.
(6) Compute the MI among all the TFs and operons by using the CLR algorithm (14).
(7) Compute the z-score among all the TFs and operons from the MI distributions by using the CLR algorithm.
(8) Infer the TFs regulating a given operon, single and cooperative interactions, according to a given

threshold depending on the desired precision. The threshold for cooperative regulations is taken higher than for single ones (2-fold for the reported calculations, although it can be modified straightforwardly) to avoid overfitting in the computation of the combinatorial interactions. See Supplementary Data for cut-off threshold selection.

(9) For each operon, estimate the kinetic parameters for its regulating TFs by using multiple linear regressions (obtaining single and synergistic interactions). Eventually, remove regulations with low strength.

(10) Construct a SBML file containing the ODE-based model using the inferred topology and the estimated kinetic parameters.

Prediction of transcriptomic profiles

To compute the performance of our algorithm, we defined a reference network taking those genes with known transcriptional regulation. In addition, the TFs that were present in our reference set regulating genes outside the reference set were also removed when determining the performance of the algorithm. Then, only the interactions among the genes present in that reference set were evaluated to compute the algorithm efficiency. All known interactions cataloged in RegulonDB version 4 (28) were used to construct the reference network in *E. coli*. However, we are still far from a complete understanding of the transcriptional regulation network of *E. coli*. Therefore, we designed *in silico* genomes with predefined regulations to validate the performance of our algorithm. For that, we did not consider: (i) operons with self-regulations; (ii) operons with constitutive promoters; and (iii) operons containing only TFs.

We calculated two types of efficiencies (precision rate and sensitivity) to compare the inferred network with the reference network. We defined precision rate as the fraction of predicted interactions that are correct $[TP/(TP+FP)]$, and sensitivity as the fraction of all known interactions that are discovered by the algorithm $[TP/(TP+FN)]$, where TP is the number of true positives, FN the number of false negatives and FP the number of false positives (39,40).

Designing genomes and expression data

In order to evaluate the suitability of our procedure to redesign the transcription regulation, we will analyze our ability to infer the kinetic parameters. Since they are not known for any organism, this lead us to the development of a Generator of Artificial Genomes (GAG) to *in silico* create expression profiles (Figure S2). To construct such genomes, we specify the number of genes and TFs (this last is usually taken one order of magnitude less than the number of genes), and eventually the ratio between inducers and repressors (we have used 2/3). We can also specify the degree of connectivity to obtain scale-free networks [we have considered a probability distribution $P(k) \propto k^{-2}$ where k is the number of regulators of an operon], and the law for clustering distribution [we have assumed $P(n) \propto 2^{-n}$ where n is the number of genes per operon]. To generate synthetic microarray data, we first obtain the steady state of the system [$y = f(y)$, since $dy/dt = f(y) - y$ with an arbitrary degradation rate of 1] without taking into account cooperations between different regulators (i.e. $\beta_{ijk} = 0, \forall i, j, k$) as an approximate solution of the system (Equation 2). In fact, as the gene expressions (y) are only functions of the TFs (y_{TF}), we can write the system as $y = f(y_{TF})$. Subsequently, we generate a new condition by randomly choosing a set of TFs with given size optimized for the inference (Figure S4) and perturbing their steady state values, while maintaining constant the other TF expressions. The perturbations over/under-express the TFs to a 50%, relative to their steady states. Hence, this perturbed value (y^*_{TF}) is used to recalculate the gene expressions by applying the model $y^* = f(y^*_{TF})$. Although this could be extended to more complicated conditions, where different gene categories are altered, the conditions based on TF perturbations are more revealing. Furthermore, to generate more realistic data we have added random fluctuations (which would simulate noisy data) in the expression values. We have studied the efficiency (precision rate and sensitivity) of our algorithm for different noise levels. In Figure S5 (see Supplementary Data) we show that InferGene maintains high efficiency up to 10% of noise amplitude.

RESULTS

Genome-wide quantitative model of *E. coli*

In the present study, we have applied inference methodologies recently used to obtain models suitable for genome redesign. We have considered the *E. coli* genome, which contains 4345 nonredundant genes, of which 328 are putative TFs. The genome is organized into 3333 operons, 2447 containing single genes and 886 polycistronic units. The reference regulatory set has been constructed according to RegulonDB (28). For the inference procedure, we have used public microarray data (41) from Affymetrix normalized using RMA (42). This is a microarray compendium containing 189 experiments. From this dataset, 20 experiments were excluded in order to later predict expression profiles from unbiased data. The inferred network contains 525 regulatory interactions (z-score > 6.92) and 566 combinatorial influences (z-score > 12). InferGene predicts 3982 genes to be controlled by constitutive promoters. In Figure 2a, a we plot the inferred transcriptional regulatory network of *E. coli* visualized using Cytoscape, having 75% of precision rate and 5% of sensitivity for single regulations, comparing with the regulations present in RegulonDB. Indeed there is a trade-off between sensitivity and precision, and the requirement of a high precision rate (such as 75%) gives very low sensitivities around 5% for *E. coli* (14). Notice that even a perfect algorithm (100% precision), where there are no false positives, could reach very low sensitivities if it is too conservative and suggest much fewer interactions than the ones in the reference set.

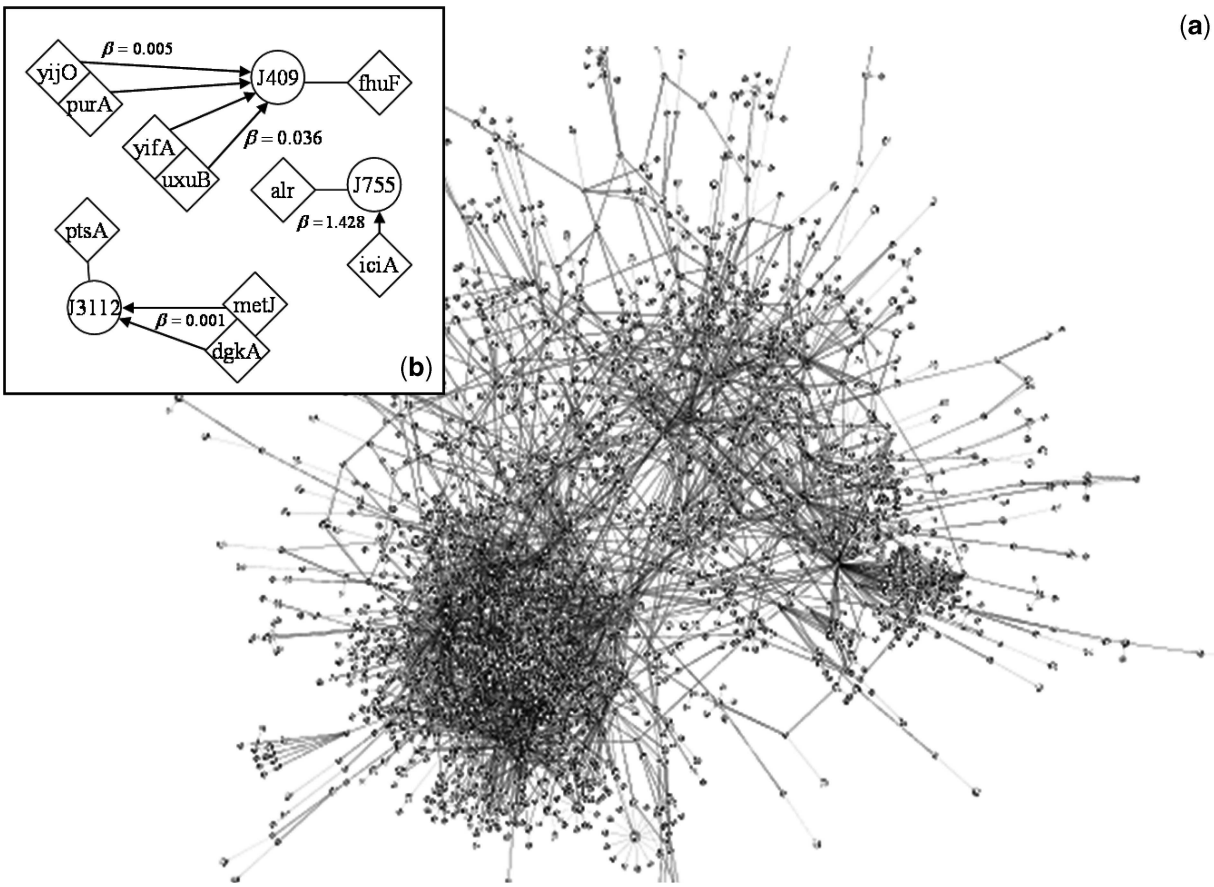

Figure 2. Inferred regulatory network of *E. coli* visualized using Cytoscape (34) thanks to the SBML import. (**a**) Full transcriptional regulatory network by InferGene with 75% of precision rate and 5% of sensitivity for single regulations (z-score > 6.92). Genes are indicated as rhombus and transcription reactions as circles. Arrows mean regulations and lines connect reactions with the corresponding gene products. We represent synergistic TFs regulations by drawing together several rhombus. The strength of each regulation can be found in the SBML model (provided in the Supplementary Data). (**b**) Example of an *E. coli* subnetwork involving genes related with the cell structure and transport (one of the best predicted biological functions, see also Figure S9). The TF *iciA* was selected by InferGene as the most likely regulator of the *alr* operon from the set of all (328) candidates. On the other hand, the TFs *metJ* and *dgkA* synergistically regulate the *ptsA* operon. Also InferGene proposes a combinatorial regulation of the *fhuF* operon: (*yijO* AND *purA*) OR (*yifA* AND *uxuB*).

To analyze those results in a biological context, we have used the EcoCyc (43) classification to group genes by biological functions and to rank those groups according to their level of prediction (see Supplementary Figure S9). We have scored each biological function as $\Delta_{bf} = \frac{1}{nm}\sum_{g\in bf}\sum_{c\in \text{set}}\left|\widehat{y}_{gc} - y_{gc}\right|$, where n is number of genes involved in the biological function, m the number of the new conditions of the set ($m = 20$), $\widehat{y}_{gc}$ the predicted expression and y_{gc} the measured expression. The best predicted functions are involved in the metabolism, such as biosynthesis of lipoprotein, carnitine, glycolate and glycoprotein, or functions related with information transfer such as rRNA and stable RNA, ATP binding, DNA and DNA degradation. In addition, we have observed two significant correlations between the number of constitutively expressed genes and the error in expression (Δ_{bf}). These genes are from biological functions involved in the location of gene products and the cell processes (see in Supplementary Figure S9). On the other hand, in Figure 2b, we show an example of such groups, where the *alr* operon, involved in metabolism of alanine biosynthesis, is regulated by *iciA* with a strength of $\beta_{iciA} = 1.428$, according to InferGene. InferGene also predicts the regulation for the *ptsA* operon, involved in the cell structure of *pilus_csgB*, where *metJ* and *dgkA* act synergistically with $\beta_{metJ,dgkA} = 0.001$. For the *fhuF* operon, involved in transport, InferGene proposes the combinatorial regulation (*yijO* AND *purA*) OR (*yjfA* AND *uxuB*), with $\beta_{yijO,purA} = 0.005$ and $\beta_{yjfA,uxuB} = 0.036$. Notice that these regulations are not found in RegulonDB, but are obtained as the best experiment-fitting regulators.

Furthermore, we provide in the Supplementary Data a list of the *E. coli* promoters classified according to their inferred regulation. An analysis of the prediction of the promoter regulation shows (see Supplementary Figure S10) that the promoters which are regulated by two TFs are better predicted. In addition, the algorithm can be used to account for nontranscriptional regulations (20). In the Supplementary Data, we have applied this to the well-known SOS pathway. There we show that an effective model of gene–gene interactions can improve

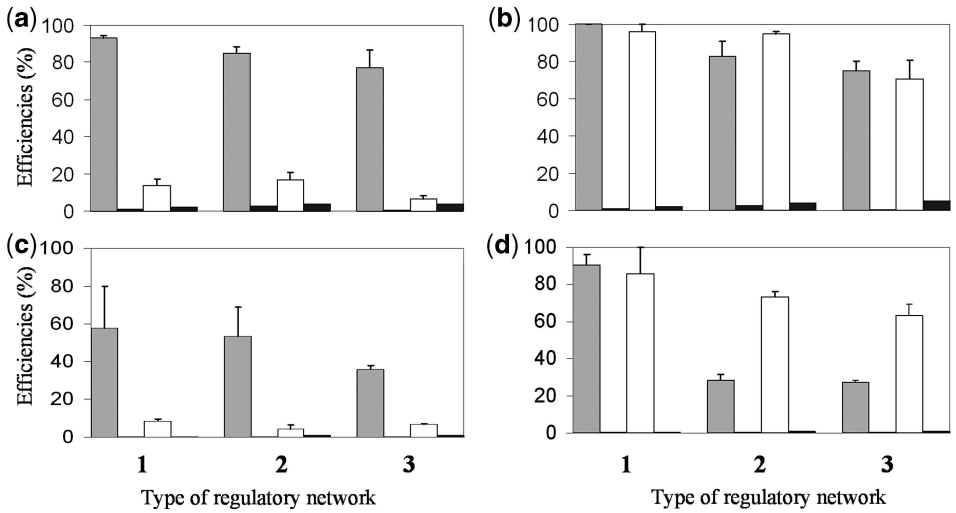

Figure 3. InferGene performance. Evaluation of sensitivity (gray) and precision rate (white) together with a random inference (black) of the transcriptional regulatory network. We used several types of synthetic genomes with different topological and parametrical properties generated by GAG. We constructed three types of genomes: (i) all promoters are regulated by at most one TF; (ii) the promoters that can be regulated by more than one TF; and (iii) promoters with combinatorial regulations including synergistic effects. Genomes for (**a,b**) had 500 genes and 50 TFs, and for (**c,d**) 5000 genes and 200 TFs. The number of conditions was in (a) 100, (b) 250, (c) 300 and (d) 600. Deviations in precision rates and sensitivities were calculated using three different genomes for each type. The *z*-score threshold used was in (a) 0.5, (b) 1, (c) 3 and (d) 7.

the prediction over the pure transcriptional one (see Figures S23–S25).

Designing genomes and validating their transcription profiles

We have constructed several genomes *in silico* using GAG and we have compared the predefined regulations in our models with the regulations inferred by InferGene. We have constructed three types of transcription networks according to the mode of regulation of its constituent operons: (i) networks with promoters regulated by at most one TF; (ii) networks with promoters that can be regulated by more than one TF; and (iii) networks with promoters that can be combinatorially regulated including synergistic effects. We have computed the precision rate and sensitivity (see Methods section) to quantify the efficiency of InferGene. In Figure 3, we show the evaluation of the inference for different types of genome networks. InferGene, which at this stage relies on CLR, predicts the 85.4% (sensitivity) of the possible interactions although only the 15.7% (precision rate) of them are correct for a genome of 500 genes using 100 conditions (Figure 3a). However, if the number of conditions increases to 250, the precision rate reaches values around the 90% (see Figure 3b). The same trend occurs with larger genomes as we can see from Figures 3c and d, where we have worked with genomes of 5000 genes with 300 and 600 conditions, respectively. Thus, we improve 6-fold the precision rate, maintaining a given level of sensitivity, when increasing the number of conditions 2.5-fold. Therefore, the efficiency of algorithm has a nonlinear behavior regarding the number of conditions used for training. We have also extended the inference capabilities of CLR to cooperative interactions. Our results show that we need a minimum set of microarray experiments to infer a transcriptional regulatory network with high precision rate for a given sensitivity. Furthermore, genomes with only promoters regulated by at most one TF reached higher values of precision rate and sensitivity.

We have analyzed the predictive power of InferGene by calculating a score based on the error made on predicting the expression levels (Δ_{op}), and other score based on the error made on the prediction of the model parameters (Γ). We define $\Delta_{op} = \frac{1}{nm}\sum_{g\in op}\sum_{c\in \text{set}}|\widehat{y}_{gc} - y_{gc}|$, where $\widehat{y}_{gc}$ is the predicted expression profile, y_{gc} is the experimental value, n is the number of operons that are correctly inferred according to RegulonDB and m is the number of conditions that were not used in the training set ($m = 20$). We also define $\Gamma = \frac{1}{nn_p}\sum_{g\in o}\sum_{p\in P}|\widehat{\beta}_{gp} - \beta_{gp}|$, where n_p is number of parameters we use to model the kinetics of the operon expression, $\widehat{\beta}_{gp}$ are the estimated model parameters and β_{gp} are the model parameters from GAG. To perform such analysis, we have generated a network using the GAG algorithm with 500 genes across 250 conditions (see Supplementary Figure S11). The median for Δ_{op} was 0.009, and for Γ was around 0.01. Moreover, we have validated the estimated parameters by performing linear regressions with the predefined kinetic models and obtaining correlations (Pearson coefficients) above 0.90 (see Supplementary Figure S3).

Prediction of wild-type *E. coli* trancriptomic profiles

Before proceeding to change the regulation of *E. coli*, we have calculated the ability of the inferred model to predict the steady state expression levels of the *E. coli* genes. For that, we have used the model together with the expression levels of all the TFs for each experimental condition to compute the global expression profile. Afterwards, we have compared the predicted expression values with the corresponding measurements, obtaining Δ_{op}. We have also determined the predictive power of the inferred

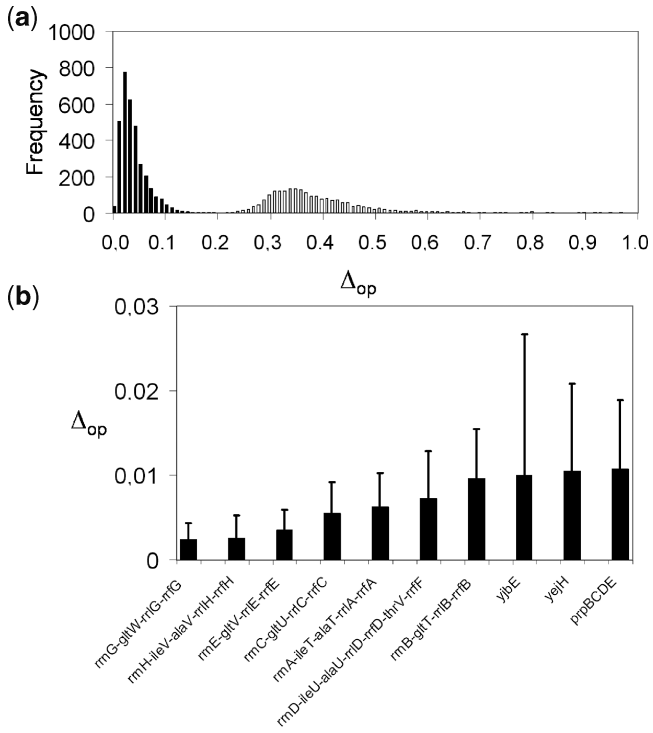

Figure 4. (**a**) Histogram of the expression error on the transcriptomic profile for each operon (Δ_{op}). In black, model with parameters from linear regression; in white, model with random parameters (for a fixed inferred topology). (**b**) We show the mean of Δ_{op} with the corresponding standard deviations for the best predicted operons. We measured the predictive power under the 20 conditions of the testing set.

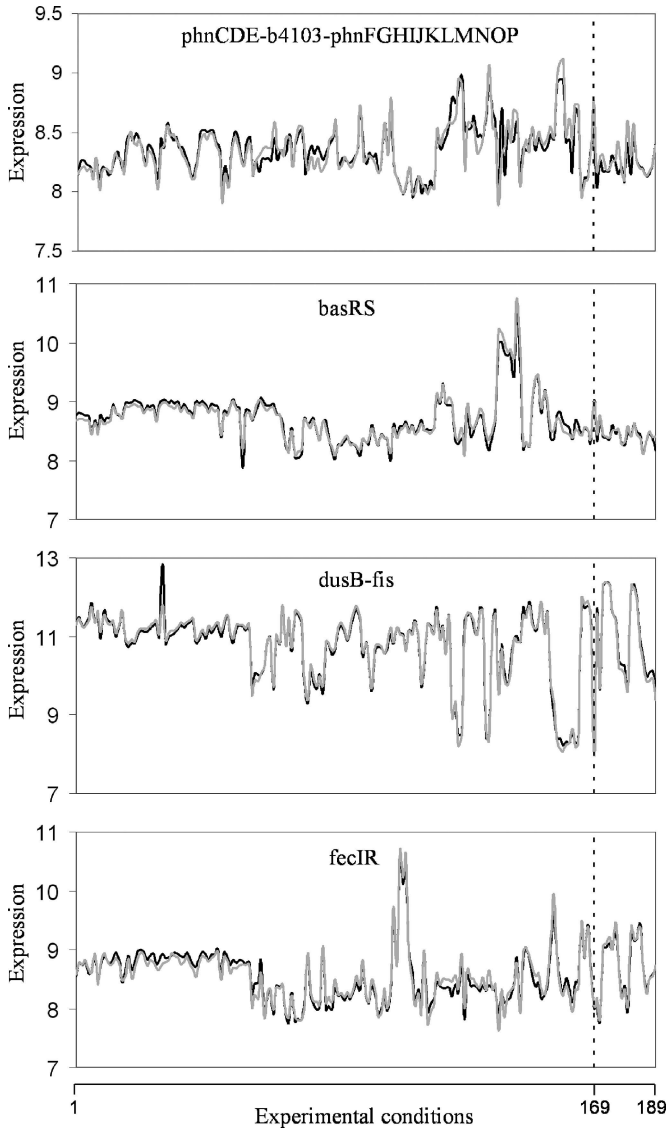

Figure 5. Prediction of expression profiles in *E. coli*. Each plot shows the experimental profile (gray line), and the profile predicted by our model (black line). The last 20 experiments, separated by a dashed line, correspond to conditions that were not included in the training dataset with which we inferred the kinetic model.

model on the 20 experimental conditions excluded from training dataset. The distribution of Δ_{op} for the 3333 operons of *E. coli* is shown in Figure 4a (black bars). The mean of this distribution is 0.048. White bars represent a model with random parameters for the inferred topology. In Figure 4b, we show the prediction for the best inferred operons. It is interesting to note that the genes from these operons are involved in functions related with information transfer (RNA related, such as transcription related, tRNA, rRNA or stable RNA; and protein related such as translation), regulation, location of gene products (cytoplasm and *ompR*) and cell processes (adaptation and defense survival).

In Figure 5, we plot the predicted profiles with lowest Δ_{op} against the experimental profiles across all conditions (189 experiments, 169 conditions from the training set and 20 new conditions for prediction). We also perform a *K*-fold cross-validation (we consider nine partitions, see Figures S13 and S14) to ensure that our results do not depend on the selection of the testing set. In the Supplementary Data, we provide the best predicted profiles for the distinct types of promoters. In addition, we have analyzed the profile prediction to evaluate the best predicted conditions (see Figure S12). We have found that the conditions upregulating genes *ruvA*, *sulA*, *umuD*, *dinP*, *recA*, *luc*, *uvrA*, *lon*, *lexA* and *dinI* are better predicted, and the experiments with plasmids pPROEx-CAT, pET3d and T7 controllable have higher error (see more details in the Supplementary Data).

Redesign of the global transcription regulation

Finally, we have used our model to predict the expression profile under knockouts of TFs (conditions from the training set). This is a first step toward changing the transcription regulation. For that, we have solved the system of equations in steady state by removing the corresponding transcription regulation. For simplicity, here we have neglected the combinatorial terms to work with a linear model and recalculated the kinetic parameters. To account for experimentally reported interactions, we have incorporated into the model regulations between pairs of TFs according to RegulonDB. In Figure 6, we plot the

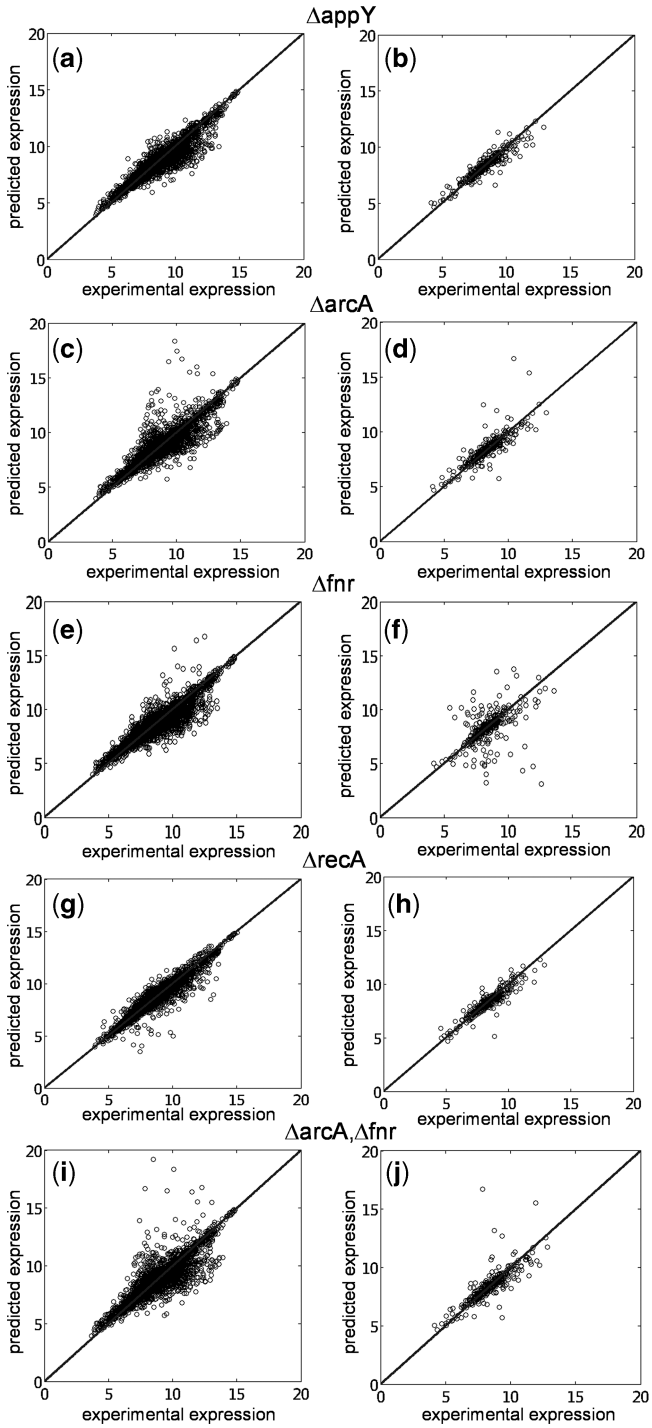

Figure 6. Prediction of expression profiles in *E. coli* from single knockouts of the TFs *appY*, *arcA*, *fnr* and *recA*, and a double knockout of the TFs *arcA* and *fnr*. In (**a, c, e, g, i**) whole transcriptome, in (**b, d, f, h, j**) TFs profile. The relative expression error (Δ_{op}) is 4% in (a), 4% in (b), 5% in (c), 5% in (d), 5% in (e), 5% in (f), 4% in (g), 4% in (h), 6% in (i) and 5% in (j). Experimental data is obtained from (41).

predicted versus the experimental profiles for the knockouts of the TFs *appY*, *arcA*, *fnr* and *recA*. In the Supplementary Data we also show predictions for the knockouts of the TFs *crp*, *cspA*, *hns*, *oxyR*, *soxS* and a double knockout of *arcA-fnr*. We show how the model is able to capture the whole transcriptomic expression due to a perturbation in the TF network (the relative expression errors, in average for all genes, are shown in Figure 6 caption). Therefore, the model quantitatively accounts for a global regulatory redesign, especially in case of knockouts of master regulators.

Moreover, we have applied our procedure to the modification of the global transcription regulation by adding new regulations into the genomic network. This was done experimentally by Isolan *et al.* (44), where they overexpressed plasmids pairing together wild-type promoters with ORFs coding for TF that were master regulators. We used our procedure to predict the gene expression of such transcriptional perturbation for the particular case where the *rpoS* and *malT* promoters are disposed together with the ORFs *ompR* and *fliA*, respectively (see Figure 7, relative expression errors are shown in the caption).

DISCUSSION

We have discussed a methodology to create quantitative models for transcription regulation aimed to future genome redesign projects. We have shown how we could use recent methodologies to infer the global topology of transcription regulation to produce the kinetic model able for genome redesign. We have successfully applied the inferred model to predict the transcriptomic response of *E. coli* under experimental conditions not included in the training set. The prediction has in average an error of 1–5% relative to the experimental value (average computed across all conditions). Furthermore, we have predicted the gene expression under knockouts of TFs and genetic rewirings (44) by solving a perturbed model, showing the predictive power of the inference procedure. Such perturbations change the regulatory map of the cell, but more complex redesigns, even a whole transcription refactorization, could be *in silico* explored by using our model. Our algorithm provides a global deterministic kinetic model of genetic regulations using microarray data. We show how to use this kinetic model to make predictions (23). Thus, our approach constitutes an important step toward the large-scale design of cell behaviors by providing models which are validated using *in silico* genomes and experimental transcription data. In this direction, we have accounted for simple transcription rewirings (44) by obtaining the gene expressions using computational methods. Such models can be used in the future to rewire the regulation of organisms without affecting their physiological behavior.

The algorithm reaches high efficiencies at the topology and kinetic level, based on the CLR algorithm (14) to infer the network together with an extension to include cooperations in combinatorial promoters. However, it could use other approaches such as Bayesian methods (19). In addition, the generation of synthetic data from specified genome models has been essential to analyze the performance and limitations of InferGene. Indeed, we have shown how the precision rate is drastically improved, from 10–20% to 80–90%, by just doubling the number

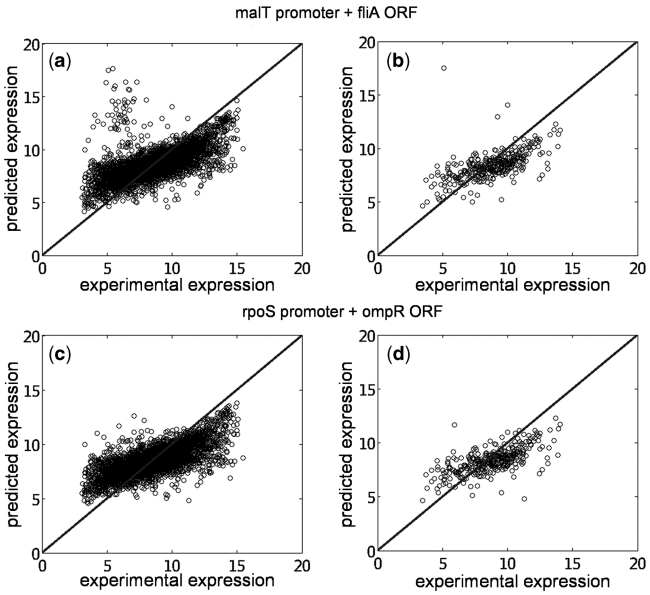

Figure 7. Prediction of expression profiles in *E. coli* from transcriptional perturbations rewiring the wild-type regulatory map putting together in a high-copy plasmid the *rpoS* and *malT* promoters with the ORFs of *ompR* and *fliA*, respectively. In (**a,c**) whole transcriptome, in (**b,d**) TFs profile. The relative expression error (Δ_{op}) is 18% in (a), 16% in (b), 19% in (c) and 16% in (d). Experimental data is obtained from (44).

of perturbations in artificial genomes. Moreover, the error in the prediction of the expression value for correctly predicted regulations is of the order of magnitude of the standard errors on measured expression data, and the estimated parameters highly correlate with the predefined ones (correlation coefficient >0.9). The inaccuracies in our prediction could be rationalized by the lack of modeling of many dynamic variables of the cell (e.g. proteins or metabolites) or nontranscriptional regulations (e.g. protein–protein or RNAi), since these variables are not experimentally measured using microarrays. Furthermore, future works could consider confidence intervals on the model parameters to analyze the stochasticity in expression data. We provide the inferred model in a standard format, as it is SBML (33), which can be used for further applications. In addition, we have used genome annotation to identify the best predicted biological functions.

Our approach can take advantage from additional sources of information. For instance, it can incorporate in the inferred model experimentally validated interactions (e.g. from functional genomics measurements or sequence analysis) as a regulatory background. In addition, the knowledge on the genome sequence can help in the inference procedure, by providing information about operon structure, identification of TFs and their regulations (28,45,46). The prior knowledge about regulation provides a topology that can be added into the model and can be used to predict new interactions with high fidelity (47). The methodology can also be applied to account for nontranscriptional interactions. In the Supplementary Data, we use the well known SOS pathway to show that an effective model of gene–gene interactions can improve the prediction over the pure transcriptional one. Furthermore, the algorithm can be expanded in a straightforward way to input expression data from time series.

The identification of regulations is a high time-consuming activity. The running time scales with the number of genes and the square of the number of conditions. Nonetheless, the parameter estimation is a quick process (relative to the previous). For instance, in *E. coli* there are 4345 genes (strain K-12) clustered in 3333 operons, and 328 TFs and 53 628 pairs of TFs (28). The whole inference process took 6 h accomplished on a computer Pentium M 2.00 GHz and 1 GB RAM (time resources for parameter estimation are neglected as they are around 2 min). However, all simulations can be run in parallel allowing the reduction of the execution time (<5 min on a simple cluster). In this way, distributed computing provides the necessary resources to apply our methodology to infer the regulations of much larger genomes. Our methodology provides a simple and fast way to obtain a quantitative global model of transcriptional regulation even for large networks. The incorporation of sparse Bayesian regression methods (19) provides a promising extension for further works. Such methods would provide better inference but increasing the computational cost.

The construction of genome-scale models is clearly a valuable step toward the understanding of the cellular behavior (4), but it is also of interest for the emerging field of synthetic biology, where functional genetic circuits are engineered into cells dealing to minimize the impact on the host (48). Hence, InferGene provides an accurate model to predict the changes in the biological processes when perturbing the cell. In addition, this model can be applied to discover molecular targets of heterologous compounds (20,21).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted with M. Elati for his careful reading of the article and his comments. We also acknowledge the anonymous reviewers for their suggestions.

FUNDING

Spanish Ministry of Education and Science (ref. TIN 2006-12860); Structural Funds of the European Regional Development Fund; EU grants BioModularH2 (FP6-NEST contract 043340) and EMERGENCE (FP6-NEST contract 043338); ATIGE Genopole/UEVE and the MIT-France grants; Graduate fellowship from the Conselleria d'Educacio de la Generalitat Valenciana (ref. BFPI 2007/160 to G.R.) and an EMBO Short-term fellowship (ref. ASTF-343.00-2007 to G.R.). HPC-Europa programme. Funding for open access charge: EU grant BioModularH2 FP6-NEST-043340.

Conflict of interest statement. None declared.

REFERENCES

1. Lee,T., Rinaldi,N., Robert,F., Odom,D., Bar-Joseph,Z., Gerber,G., Hannett,N., Harbison,C., Thompson,C., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
2. deJong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comp. Biol.*, **9**, 67–103.
3. Hughes,T., Marton,M., Jones,A., Roberts,C., Stoughton,R., Armour,C., Bennett,H., Coffey,E., Dai,H., He,Y. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
4. Covert,M.W., Knight,E.M., Reed,J.L., Herrgard,M.J. and Palsson,B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.
5. Eisen,M., Spellman,P., Brown,P. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
6. Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.
7. Alon,U., Barkai,N., Notterman,D., Gish,K., Ybarra,S., Mack,D. and Levine,A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.
8. Dhaeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.
9. Ihmels,J., Friedlander,G., Bergmann,S., Sarig,O., Ziv,Y. and Barkai,N. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.
10. Bansal,M., Belcastro,V., Ambesi-Impiombato,A. and diBernardo,D. (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**, 78.
11. Butte,A. and Kohane,I. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropymeasurements. *Pac. Symp. Biocomp.*, **5**, 415–426.
12. Basso,K., Margolin,A. A., Stolovitzky,G., Klein,U., Dalla-Favera,R. and Califano,A. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.
13. Margollin,A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., dellaFavera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
14. Faith,J., Hayete,B., Thaden,J., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J. and Gardner,T. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *Plos Biol.*, **5**, e8.
15. Meyer,P.E., Kontos,K., Lafitte,F. and Bontempi,G. (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinf. Syst. Biol.*, **2007**, 79879.
16. Yu,J., Smith,V., Wang,P., Hartemink,A. and Jarvis,E. (2004) Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.
17. Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.
18. Fujita,A., Sato,J.R., Garay-Malpartida,H.M., Yamaguchi,R., Miyano,S., Sogayar,M.C. and Ferreira,C.E. (2007) Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Syst. Biol.*, **1**, 39.
19. Steinke,F., Seeger,M. and Tsuda,K. (2007) Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. *BMC Syst. Biol.*, **1**, 51.
20. Gardner,T., diBernardo,D., Lorenz,D. and Collins,J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiles. *Science*, **301**, 102–105.
21. diBernardo,D., Thompson,M., Gardner,T., Chobot,S., Eastwood,E., Wojtovich,A., Elliott,S., Schaus,S. and Collins,J. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, **3**, 377–383.
22. Shevade,S. and Keerthi,S. (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, **19**, 2246–2253.
23. Bonneau,R., Reiss,D., Shannon,P., Facciotti,M., Hood,L., Baliga,N. and Thorsson,V. (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.*, **7**, R36.
24. Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B.*, **58**, 267–288.
25. Behrens,J., vonKries,J., Khl,M., Bruhn,L., Wedlich,D., Grosschedl,R. and Birchmeier,W. (1996) Functional interaction of bold β-catenin with the transcription factor LEF-1. *Nature*, **328**, 638–642.
26. Stewart,V. and Bledsoe,P. (2005) Fnr-, NarP- and Narl-dependent regulation of transcription initiation from the *Haemophilus influenzae* Rd napF (Periplasmic Nitrate Reductase) promoter in *Escherichia coli* K-12. *J. Bacteriol.*, **187**, 6928–6935.
27. Long,J. and Roth,M. (2008) Synthetic microarray data generation with RANGE and NEMO. *Bioinformatics*, **24**, 132–134.
28. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J. *et al.* (2006) Regu-lonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **34**, D394.
29. Gray,R. (1990) *Entropy and Information Theory*. Springer-Verlag, New York, NY, USA.
30. Steuer,R., Kurths,J., Daub,C.O., Weise,J. and Selbig,J. (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, S231–S240.
31. Daub,C., Steuer,R., Selbig,J. and Kloska,S. (2004) Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, **5**, 118.
32. Cohen,J.P.C., West,S. and Aiken,L. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.
33. Hucka,M., Bolouri,H., Finney,A., Sauro,H., Doyle,J.K.H., Arkin,A., Bornstein,B., Bray,D., Cornish-Bowden,A., Cuellar,A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
34. Shannon,P., Markiel,A., Ozier,O., Baliga,N., Wang,J., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
35. Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.
36. Affymetrix (1999) *Affymetrix Microarray Suite User Guide, version 4*. Affymetrix, Santa Clara, CA, USA.
37. Sabatti,C., Rohlin,L., Oh,M. and Liao,J. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **30**, 2886–2893.
38. Dongarra,J., Bunch,J., Moler,C. and Stewart,P. (1979) *LINPACK User's Guide*. SIAM, Philadelphia, PA, USA.
39. Altman,D. and Bland,J. (1994) Statistics notes: diagnostic tests 1: sensitivity and specificity. *Br. Med. J.*, **308**, 1552.
40. Altman,D. and Bland,J. (1994) Statistics notes: diagnostic tests 2: predictive values. *Br. Med. J.*, **309**, 102.
41. Faith,J., Driscoll,M., Fusaro,V., Cosgrove,E., Hayete,B., Juhn,F., Schneider,S. and Gardner,T. (2008) Many microbe microarrays database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.
42. Irizarry,R., Hobbs,B., Collin,F., Beazer-Barclay,Y., Antonellis,K., Scherf,U. and Speed,T. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
43. Karp,P., Riley,M., Saier,M., Paulsen,I., Collado-Vides,J., Paley,S., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc DataBase. *Nucleic Acids Res.*, **30**, 56–58.

44. Isalan,M., Lemerle,C., Michalodimitrakis,K., Horn,C., Beltrao,P., Raineri,E., Garriga-Canut,M. and Serrano,L. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, **452**, 840–845.
45. Price,M., Huang,K., Alm,E. and Arkin,A. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
46. Reiss,D., Baliga,N. and Bonneau,R. (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **7**, 280.
47. Mordelet,F. and Vert,J.-P. (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **24**, i76–i82.
48. Sprinzak,D. and Elowitz,M. (2005) Reconstruction of genetic circuits. *Nature*, **438**, 443–448.